Unit 4: Basic Statistical Concepts
IB Math AA SL

Answer all 15 questions. Show all working. For Paper 1 questions, use analytical methods. For Paper 2 questions, use your graphic display calculator (GDC) efficiently.

1. **[Paper 1 Style, Non-Calculator, Easy, 4 marks]**
   State whether the following sets of data are discrete or continuous:

   (a) The exact time taken by an athlete to complete a 400 m sprint.

   (b) The number of red cars parked in a school car park.

   (c) The shoe sizes of students in a mathematics class.

   (d) The exact volume of water remaining in a water bottle.

2. **[Paper 2 Style, Calculator Required, Easy, 4 marks]**
   A researcher is conducting a survey. Match the following scenarios to the most appropriate sampling technique from the list: *Simple Random, Systematic, Stratified, Convenience, Quota.*

   (a) The researcher stands outside a supermarket and interviews the first 50 people who walk past.

   (b) The researcher selects a random starting point on a list of residents and then selects every 15^{th} person on the list.

   (c) The researcher needs 30 men and 30 women. He stands on the street and interviews people until he has exactly 30 of each, actively seeking out specific genders to meet his numbers.

   (d) The researcher numbers a population from 1 to 500 and uses a random number generator on their GDC to select 20 unique numbers.

3. **[Paper 2 Style, Calculator Required, Easy, 5 marks]**
The weights, in kilograms, of eight dogs visiting a veterinary clinic are recorded as follows:

$$12, \quad 15, \quad 16, \quad 18, \quad 19, \quad 21, \quad 22, \quad 35$$

An outlier is defined mathematically as any value that is greater than $Q_3 + 1.5 \times \text{IQR}$ or less than $Q_1 - 1.5 \times \text{IQR}$.

(a) Find the lower quartile (Q_1) and the upper quartile (Q_3).

(b) Calculate the interquartile range (IQR).

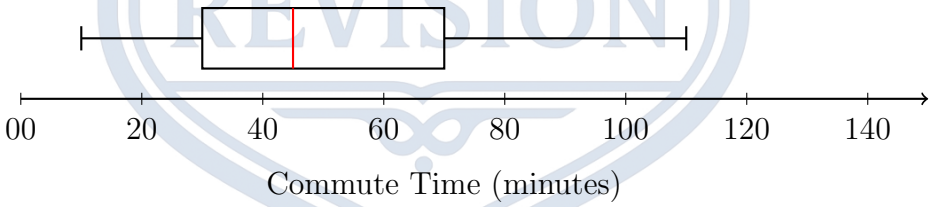
(c) Determine mathematically whether the weight of 35 kg is an outlier.

4. **[Paper 2 Style, Calculator Required, Easy, 4 marks]**
A school has 850 students. The principal wants to select a simple random sample of 12 students to participate in a focus group about school uniform.

(a) Define the term *population* in the context of this scenario.

(b) Describe a method the principal could use to select this simple random sample using their graphic display calculator.

5. **[Paper 2 Style, Calculator Required, Medium, 5 marks]**
The daily commute times (in minutes) for 60 workers are summarized in the box-and-whisker diagram below.



(a) State the median commute time.

(b) Calculate the outlier boundaries.

(c) Given that the maximum value shown is 110 minutes, does this dataset contain any outliers? Justify your answer.

6. **[Paper 2 Style, Calculator Required, Medium, 4 marks]**
A multinational company wants to survey its employees regarding a new remote working policy. The company consists of three main departments: Sales (120 employees), IT (45 employees), and Human Resources (35 employees). The HR manager decides to take a stratified sample of 40 employees.

   (a) Explain why a stratified sample is more appropriate than a simple random sample in this context.

   (b) Calculate the exact number of employees that should be sampled from the IT department.

7. **[Paper 1 Style, Non-Calculator, Medium, 4 marks]**
A fitness magazine wants to estimate the average number of hours adults in a city spend exercising each week. They conduct a poll on their own website and receive 2000 responses.

   (a) Identify the target population and the sample in this scenario.

   (b) State one type of bias that is likely to be present in this sampling method, and explain how it would affect the reliability of the data source.

8. **[Paper 2 Style, Calculator Required, Medium, 4 marks]**
A university has a roster of 1200 students listed alphabetically. A researcher wants to conduct a systematic sample of 50 students.

   (a) Calculate the step interval (k) that the researcher should use.

   (b) The researcher randomly selects the number 14 as the starting point. Write down the position number on the list of the 2^{nd}, 3^{rd}, and 50^{th} students selected.

9. **[Paper 2 Style, Calculator Required, Medium, 5 marks]**
The average score of 10 students on a statistics quiz is 65. While reviewing the data, the teacher realizes there was an error in recording the data. One student's score was recorded as 40, but their actual score was 85.

   (a) Calculate the correct mean score of the 10 students.

   (b) Explain whether the median score will necessarily change as a result of correcting this error.

10. **[Paper 2 Style, Calculator Required, Hard, 4 marks]**
An environmental agency wants to sample 100 rivers across a country to test for pollution. They decide to divide the country into 4 geographic regions. They are debating between using a *Stratified Sample* and a *Quota Sample*.

    (a) State the primary difference in how the actual rivers are selected within each geographic group for these two methods.

    (b) Explain why the Stratified Sample would generally be considered a more reliable data source than the Quota Sample.

11. **[Paper 1 Style, Non-Calculator, Hard, 5 marks]**
Consider an ordered dataset of six integers:

$$5, \quad 8, \quad 9, \quad 12, \quad 15, \quad k$$

where $k > 15$. Find the least possible integer value of k such that k is classified as a mathematical outlier. Show all algebraic steps clearly.

12. **[Paper 2 Style, Calculator Required, Hard, 5 marks]**
A chemistry student measures the temperature of a reacting liquid, in degrees Celsius, at random intervals. The six recorded temperatures are:

$$20.1, \quad 22.4, \quad 25.0, \quad 26.9, \quad 85.2, \quad 31.0$$

    (a) Calculate the mean and median of these six temperatures.

    (b) Identify the anomalous data point. Suggest a realistic reason for this error in the recording of the data.

    (c) State which measure of central tendency (mean or median) is more heavily affected by this error, and calculate its new value when the anomaly is removed.

13. **[Paper 1 Style, Non-Calculator, Very Hard, 6 marks]**
A dataset consists of n completely identical values, x.

    (a) Write down the interquartile range (IQR) of this dataset.

    (b) A new, single data value y is added to the dataset, where $y \neq x$. Assuming n is very large, prove algebraically why the value y will always be classified as an outlier, regardless of how close y is to x.

    (c) Discuss the context of this mathematical quirk: why does the standard $1.5 \times$ IQR rule fail to provide useful real-world analysis for datasets with extremely low variance?

14. [**Paper 2 Style, Calculator Required, Very Hard, 6 marks**]
A large hospital employs 100 doctors and 900 nurses. An external auditor takes a simple random sample of 20 staff members to interview about working conditions.

(a) Using a binomial probability model $X \sim B(n, p)$, find the probability that the auditor randomly selects exactly 2 doctors in their sample of 20.

(b) The auditor decides to change their methodology to a stratified sample of 20 staff members. State the number of doctors that will be selected under this new method.

(c) Explain why the stratified sampling technique eliminates the sampling variation demonstrated in part (a), making it a more reliable sampling technique for representing the hospital's workforce.

15. [**Paper 2 Style, Calculator Required, Very Hard, 7 marks**]
The time taken, t in seconds, for 12 athletes to complete a 100 m sprint is recorded below:

10.1, 10.3, 10.5, 10.6, 10.6, 10.8, 11.0, 11.1, 11.3, 11.5, 11.9, 15.2

(a) Enter this data into your GDC to find the upper quartile (Q_3) and the lower quartile (Q_1).

(b) Demonstrate mathematically that the time of 15.2 s is an outlier.

(c) A sports statistician argues that although 15.2 s is a mathematical outlier, it should **not** be removed from the dataset before publishing the average sprint time. Give a contextual reason that justifies the statistician's decision to keep the outlier.